A.5 Heterogeneous Ensemble Classifiers

Authors: Danny Dunlavy (1415), Sean Gilpin (1415)

Introduction

Recent results in solving classification problems indicate that the use of ensembles classifier models often leads to improved performance over using single classifier models [1, 2, 3, 4]. In this talk, we discuss heterogeneous ensemble classifier models, where the member classifier models are not of the same model type. A discussion of the issues associated with creating such classifiers along with a brief description of the new HEterogeneous Machine Learning Open Classification Kit (HEMLOCK) will be presented. Results for a problem of text classification and several standard multi-class test problems illustrate the performance of heterogeneous ensemble classifiers.

Heterogeneous Ensemble Classifiers

Classification is the task of learning a target function that maps data instances to one of several predefined categories. These target functions are also called classifiers, classifier models, and hypotheses. We refer to a classifier constructed or learned from an ensemble of different types of classifiers is as a *heterogeneous ensemble classifier*. Note that such classifier models are also referred to as hybrid ensemble classifiers.

There are several challenges associated with learning heterogeneous ensemble classifiers. First, the choice of which base classifiers (i.e., ensemble member classifier models) needs to be determined. Performance of classifiers differs across different data sets, and thus choosing the collection of classifiers that will best classify a given set of data is often a difficult task. Each base classifier can be parametrized in many different ways, and thus an understanding of how these parameters are correlated within each base classifier as well as across the ensemble is key to classifying data sets accurately.

Fusion and Selection

A further challenge is combining base classifiers effectively, so that the performance of the ensemble classifier is better than that of the individual classifiers. There are two basic strategies for combining classifiers in an ensemble: fusion and selection [5]. Ensembles that use selection try to find the best classifier ensemble member that is most capable of correctly classifying a particular instance. Ensembles that use selection are also known as cooperative ensembles. In contrast to selection, fusion methods make use of the outputs of all of the classifiers to try determine the label of an instance. Voting is an example of fusion: each of the classifiers in the ensemble is given one vote and all of the votes are counted towards deciding which output label should be chosen. Ensembles that use fusion are commonly referred to as competitive ensembles. There are three levels at which classifiers output can be combined using fusion: label, ranking, measurement. At the label level the ensemble will only use the one class label that each of the base classifiers determines is correct. For ranking, base classifiers in the ensemble provide a ranked list of class labels reflecting how likely

each class is marked as the correct label for each data instance. Finally, at the measurement level each of the base classifiers provides output that is intrinsic to the particular learning algorithm used. Typically, measurements consist of probability distributions of the class assignment for each instance.

Diversity

It has been shown that the strength of an ensemble is related to the performance of the base classifiers and the lack of correlation between them (i.e., model diversity) [3, 4]. One way to decrease the correlations between the classifiers while increasing or maintaining the overall performance of the ensemble classifier is to include base classifiers derived from different learning algorithms such as decision trees, neural networks, perceptrons, support vector machine, etc.

HEMLOCK

HEMLOCK is a new software tool for constructing, evaluating, and applying heterogeneous ensemble data models for use in solving classification problems involving data with continuous or discrete features. HEMLOCK consists of various data readers, machine learning algorithms, model combination and comparison routines, evaluation methods for model performance testing, and interfaces to external, state-of-the-art machine learning software libraries.HEMLOCK uses XML for all input and output, and standard readers and writers are being used for data input and output. Data models are created by a variety of supervised learning methods: decision tree and random forest inducers plus a linear perceptron learner as part of HEMLOCK along with interfaces to the methods available in the WEKA software library of machine learning algorithms. Evaluation methods for assessing individual model performance include accuracy computation, confusion matrix generation, receiver operating characteristics (ROC) analysis, and area under the curve (AUC) analysis. Methods for combining heterogeneous models into a single ensemble model include majority voting and parameter regression.

Applications

In this workshop, two applications of heterogeneous ensemble classification will be discussed: e-mail classification [2] and image classification. For the e-mail classification problem, a heterogeneous ensemble of random forest, naive Bayes, and perceptron classifiers were used as base classifiers (both as individual classifiers and in homogeneous ensembles). The image classification problem consisted of labeling the number represented in handwritten digits, and heterogeneous ensembles were created using the new HEMLOCK software framework.

References

[1] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, *A comparison of decision tree ensemble creation techniques*, IEEE Trans. Pat. Recog. Mach. Int., 29 (2007), pp. 173–180.

[2] J. D. Basilico, D. M. Dunlavy, S. J. Verzi, T. L. Bauer, and W. Shaneyfelt, *Yucca mountain LSN archive assistant*, Technical report, SAND2008-1622, Sandia National Laboratories, 2008.

[3] S. Bian and W. Wang, *On diversity and accuracy of homogeneous and heterogeneous ensembles*, Intl. J. Hybrid Intel. Sys., 4 (2007), pp. 103–128.

[4] W. Wang, D. Partridge, and J. Etherington, *Hybrid ensembles and coincident failure diversity*, in Proc. International Joint Conference on Neural Networks, 2001.

[5] K. Woods, K. Bowyer, and W. P. Kegelmeyer, *Combination of multiple classifiers using local accuracy estimates*, IEEE Trans. Pat. Recog. Mach. Int., 19 (1997), pp. 405–410.

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.